Bayesian Nonparametrics

Part I

Peter Orbanz

OVERVIEW

Today

1. Basic terminology
2. Clustering
3. Latent feature models

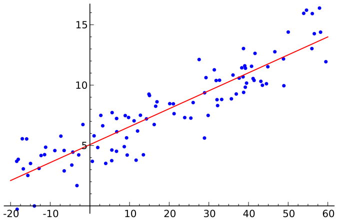
Tomorrow

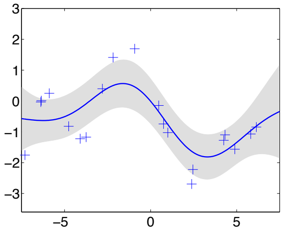
5. Constructing nonparametric Bayesian models
6. Exchangeability
7. Asymptotics

Parameters and Patterns

Parameters

$$P(X|\theta) \quad = \quad \text{Probability[data|pattern]}$$





Inference idea

$$\text{data} \quad = \quad \text{underlying pattern} \quad + \quad \text{independent noise}$$

TERMINOLOGY

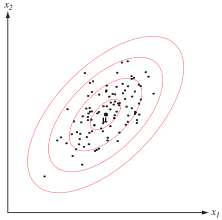
Parametric model

- Number of parameters fixed (or constantly bounded) w.r.t. sample size

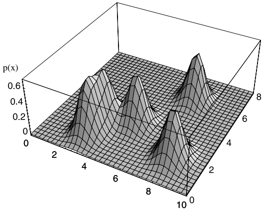
Nonparametric model

- Number of parameters grows with sample size
- ∞-dimensional parameter space

Example: Density estimation



Parametric



Nonparametric

Definition

A nonparametric Bayesian model is a Bayesian model on an ∞-dimensional parameter space.

Interpretation

Parameter space $\mathcal{T}$ = set of possible patterns, for example:

Problem	$\mathcal{T}$
Density estimation	Probability distributions
Regression	Smooth functions
Clustering	Partitions

Solution to Bayesian problem = posterior distribution on patterns

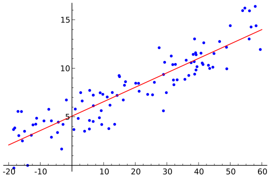
Can we justify our assumptions?

Recall:

$$\text{data} = \text{pattern} + \text{noise}$$

In Bayes' theorem:

$$Q(d\theta|x_1,\ldots,x_n) = \frac{\prod_{j=1}^n p(x_j|\theta)}{p(x_1,\ldots,x_n)} Q(d\theta)$$



Definition

$X_1, X_2, \ldots$ are *exchangeable* if $P(X_1, X_2, \ldots)$ is invariant under any permutation σ:

$$P(X_1 = x_1, X_2 = x_2, \ldots) = P(X_1 = x_{\sigma(1)}, X_2 = x_{\sigma(2)}, \ldots)$$

In words:

Order of observations does not matter.

De Finetti's Theorem

$$P(X_1 = x_1, X_2 = x_2, \ldots) = \int_{\mathbf{M}(\mathcal{X})} \Big(\prod_{j=1}^{\infty} \theta(X_j = x_j) \Big) Q(d\theta)$$

$$\Updownarrow$$

$$X_1, X_2, \ldots \text{ exchangeable}$$

where:

- $\mathbf{M}(\mathcal{X})$ is the set of probability measures on $\mathcal{X}$
- θ are values of a random probability measure Θ with distribution Q

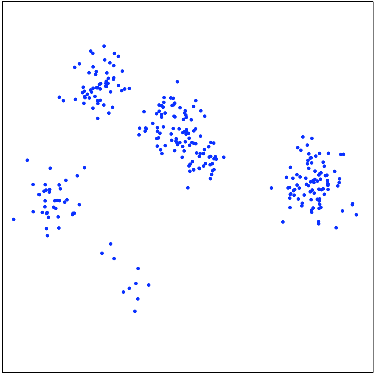
Implications

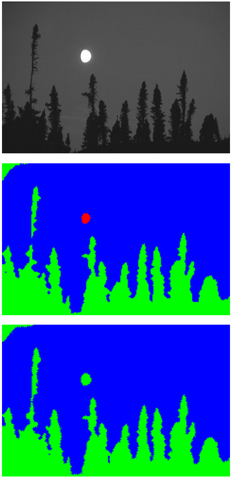
- Exchangeable data decomposes into pattern and noise
- More general than i.i.d.-assumption
- Caution: θ is in general an ∞-dimensional quantity

CLUSTERING

CLUSTERING

- Observations $X_1, X_2, \ldots$
- Each observation belongs to exactly one cluster
- Unknown pattern = partition of $\{1, \ldots, n\}$ or $\mathbb{N}$





Mixture models

$$p(x|m) = \int_{\Omega_\theta} p(x|\theta)m(d\theta)$$

m is called the *mixing measure*

Two-stage sampling

Sample $X \sim p(\,.\,|m)$ as:

1. $\Theta \sim m$
2. $X \sim p(\,.\,|\theta)$

Finite mixture model

$$p(x|\boldsymbol{\theta},\mathbf{c}) = \int_{\Omega_\theta} p(x|\theta)m(d\theta) \qquad \text{with} \qquad m(\,.\,) = \sum_{k=1}^{K} c_k \delta_{\theta_k}(\,.\,)$$

Bayesian MM

Random mixing measure

$$M(\,.\,) = \sum_{k=1}^{K} C_k \delta_{\Theta_k}(\,.\,)$$

Conjugate priors

A Bayesian model is *conjugate* if the posterior is an element of the same class of distributions as the prior ("closure under sampling").

$p(x\|\theta)$	conjugate prior
$\frac{1}{Z(\theta)} h(x) \exp(\langle S(x), \theta \rangle)$	$\frac{1}{K(\lambda, y)} \exp(\langle \theta, y \rangle - \lambda \log Z(\theta))$
Gaussian	Gaussian/inverse Wishart
multinomial	Dirichlet
...	...

Choice of priors in BMM

- Choose conjugate prior for each parameter
- In particular: Dirichlet prior on $(C_1, \dots, C_k)$

Dirichlet process

A Dirichlet process is a distribution on random probability measures of the form

$$M(\,.\,) = \sum_{k=1}^{\infty} C_k \delta_{\Theta_k}(\,.\,) \qquad \text{where} \qquad \sum_{k=1}^{\infty} C_k = 1$$

Constructive definition of DP (α, G_0)

$$\begin{aligned} \Theta_k &\sim_{\text{iid}} G_0 \\ V_k &\sim_{\text{iid}} \text{Beta}(1, \alpha) \end{aligned}$$

Compute C_k as

$$C_k := V_k \prod_{i=1}^{k-1}(1 - V_i)$$

"Stick-breaking construction"

POSTERIOR DISTRIBUTION

DP Posterior

$$\theta_{n+1}|\theta_1,\ldots,\theta_n \sim \frac{1}{n+\alpha}\sum_{j=1}^{n}\delta_{\theta_j}(\theta_{n+1}) + \frac{\alpha}{n+\alpha}G_0(\theta_{n+1})$$

Mixture Posterior

$$p(x_{n+1}|x_1,\ldots,x_n) = \sum_{k=1}^{K_n}\frac{n_k}{n+\alpha}p(x_{n+1}|\theta_k^*) + \frac{\alpha}{n+\alpha}\int p(x_{n+1}|\theta)G_0(\theta)d\theta$$

Conjugacy

- The posterior of DP (α, G_0) is DP $\left(\alpha+n, \frac{1}{n+\alpha}(\sum_k n_k\delta_{\theta_k^*} + \alpha G_0)\right)$
- Hence: The Dirichlet process is conjugate.

Latent variables

$$p(x_{n+1}|x_1,\dots,x_n) = \sum_{k=1}^{K_n} \frac{n_k}{n+\alpha} p(x_{n+1}|\theta_k^*) + \frac{\alpha}{n+\alpha} \int p(x_{n+1}|\theta) G_0(\theta) d\theta$$

We do not actually observe the Θ_j (they are latent). We observe X_j.

Assignment probabilities

$$\begin{pmatrix} q_{10} & q_{11} & \dots & q_{1K_n} \\ \vdots & \vdots & & \vdots \\ q_{n0} & q_{n1} & \dots & q_{nK_n} \end{pmatrix}$$

Where:

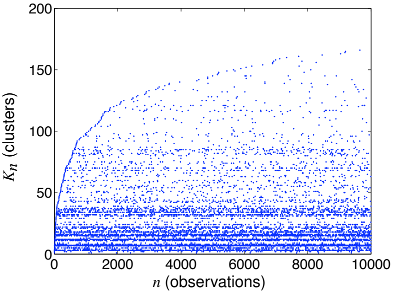
- $q_{jk} \propto n_k p(x_j|\theta_k^*)$
- $q_{j0} \propto \alpha \int p(x_j|\theta) G_0(\theta) d\theta$

Gibbs Sampling

Uses an assignment variable ϕ_j for each observation X_j.

- Assignment step: Sample $\phi_j \sim \text{Multinomial}(q_{j0},\dots,q_{jK_n})$
- Parameter sampling: $\theta_k^* \sim G_0(\theta_k^*) \prod_{x_j \in \text{ Cluster } k} p(x_j|\theta_k^*)$

Dirichlet process



$K_n = \#$ clusters in sample of size n

$$\mathbb{E}[K_n] = O(\log(n))$$

Modeling assumption

- Parametric clustering: K_∞ is *finite* (possibly unknown, but fixed).
- Nonparametric clustering: K_∞ is *infinite*

Rephrasing the question

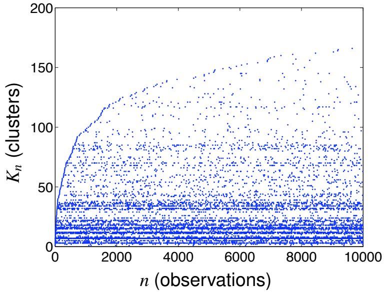
- Estimate of K_n is controlled by distribution of the cluster sizes C_k in $\sum_k C_k \delta_{\Theta_k}$.
- Ask instead: What should we assume about the distribution of C_k?

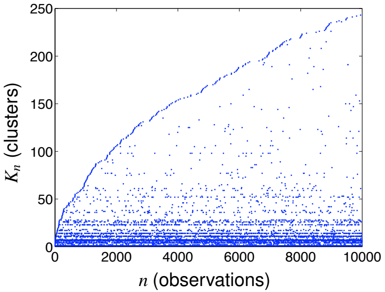
Pitman-Yor process

$$p(x_{n+1}|x_1,\ldots,x_n) = \sum_{k=1}^{K_n} \frac{n_k - d}{n+\alpha} p(x_{n+1}|\theta^*_k) + \frac{\alpha + K_n \cdot d}{n+\alpha} \int p(x_{n+1}|\theta) G_0(\theta) d\theta$$

Discount parameter $d \in [0, 1]$.

Cluster sizes





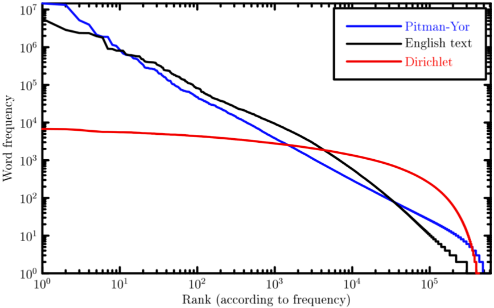
The distribution of cluster sizes is called a *power law* if

$$C_j \sim \gamma(\beta) \cdot j^{-\beta} \qquad \text{for some } \beta \in [0,1] \, .$$

Examples of power laws

- Word frequencies
- Popularity (number of friends) in social networks

Pitman-Yor language model



Discrete measures and partitions

Sampling from a discrete measure determines a *partition* of $\mathbb{N}$ into blocks b_k:

$$\Theta_n \sim_{\text{iid}} \sum_{k=1}^{\infty} c_k \delta_{\theta_k^*} \qquad \text{and set} \qquad n \in b_k \quad \Leftrightarrow \quad \Theta_n = \theta_k^*$$

As $n \longrightarrow \infty$, the block proportions converge: $\frac{|b_k|}{n} \longrightarrow c_k$

Induced random partition

The distribution of a random discrete measure $M = \sum_{k=1}^{\infty} C_k \delta_{\Theta_k}$ induces the distribution of a *random partition* $\Pi = (B_1, B_2, \dots)$.

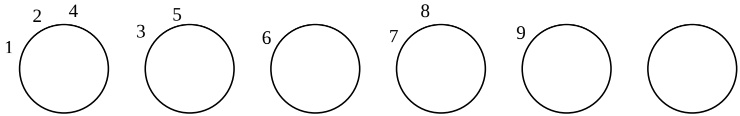
Exchangeable random partitions

- Π is called *exchangeable* if its distribution depends only on the sizes of its blocks.
- All exchangeable random paritions, and only those, can be represented by a random discrete distribution as above (Kingman's theorem).

Chinese Restaurant Process

The distribution of the random partition induced by the Dirichlet process is called the *Chinese Restaurant Process*.

"Customers and tables" analogy

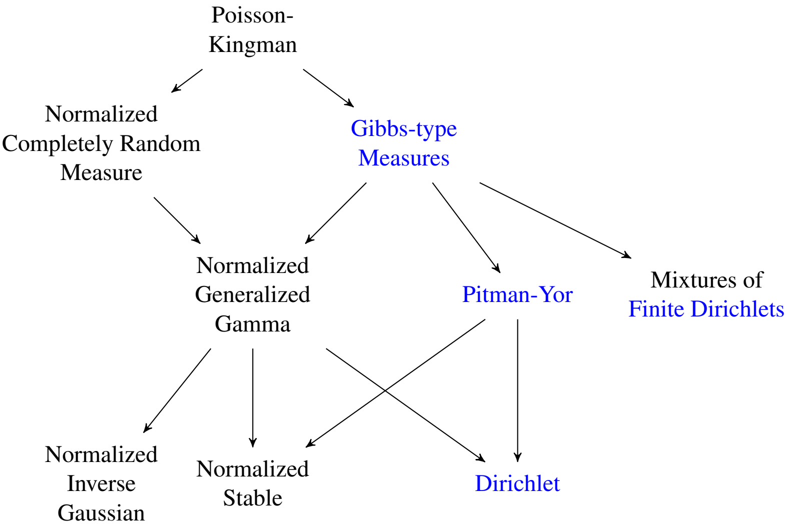


Customers = observations (indices in $\mathbb{N}$)
Tables = clusters (blocks)

Historical remark

- Originally introduced by Dubins & Pitman as a distribution on infinite permutations
- A permutation of n items defines a partition of $\{1, \ldots, n\}$ (regard cycles of permutation as blocks of partition)
- The induced distribution on partitions is the CRP we use in clustering

Families of Exchangeable Random Partitions



Random Discrete Measures

Classification (due to Prünster)

class	probability of new cluster	prior class
I	$\mathbb{P}\{\Theta_{n+1} \in \text{new cluster} \| \Theta^{(n)}\} = f(n)$	Dirichlet processes
II	$\mathbb{P}\{\Theta_{n+1} \in \text{new cluster} \| \Theta^{(n)}\} = f(n, K_n)$	Gibbs-type measures
III	$\mathbb{P}\{\Theta_{n+1} \in \text{new cluster} \| \Theta^{(n)}\} = f(n, K_n, \mathbf{n})$	

General partition priors

- Gibbs-type measures are completely classified [GP06b]
- Properties of some cases well-studied, e.g.:
  - Dirichlet process
  - Pitman-Yor process
  - Normalized inverse Gaussian process [LMP05b]
- In the future: We will have a range of models which express different prior assumptions on the distribution of cluster sizes.

Summary: Clustering

Nonparametric Bayesian clustering

- Infinite number of clusters, $K_n \leq n$ of which are observed.
- If partition exchangeable, it can be represented by a random discrete distribution.

Inference

Latent variable algorithms, since assignments ($\equiv$ partition) not observed.

- Gibbs sampling
- Variational algorithms

Prior assumption

- Distribution of cluster sizes.
- Implies prior assumption on number K_n of clusters.

Latent Feature Models

INDIAN BUFFET PROCESS

Latent feature models

- Grouping problem with overlapping clusters.
- Encode as binary matrix: Observation n in cluster k $\quad\Leftrightarrow\quad X_{nk} = 1$
- Alternatively: Item n possesses feature k $\quad\Leftrightarrow\quad X_{nk} = 1$

Indian buffet process (IBP)

1. Customer 1 tries Poisson(α) dishes.
2. Subsequent customer $n + 1$:
   - tries a previously tried dish k with probability $\dfrac{n_k}{n+1}$,
   - tries Poisson $\left(\dfrac{\alpha}{n+1}\right)$ new dishes.

Properties

- An exchangeable distribution over finite sets (of dishes).
- Intepretation:
  Observation (= customer) n in cluster (= dish) k if customer "tries dish k"

Alternative description

1. Sample $w_1, \dots, w_K \sim_{\text{iid}} \text{Beta}(1, \alpha/K)$
2. Sample $X_{1k}, \dots, X_{nk} \sim_{\text{iid}} \text{Bernoulli}(w_k)$

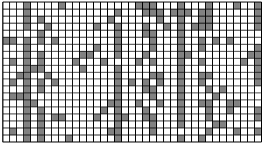
$$\begin{array}{ccc} w_1 & \dots & w_K \\ \end{array}$$
$$\begin{pmatrix} X_{11} & \dots & X_{1K} \\ \vdots & & \vdots \\ X_{N1} & \dots & X_{NK} \end{pmatrix}$$

We need some form of limit object for $\text{Beta}(1, \alpha/K)$ for $K \to \infty$.

Beta Process (BP)

Distribution on objects of the form

$$\theta = \sum_{k=1}^{\infty} w_k \delta_{\phi_k} \qquad \text{with } w_k \in [0,1] \; .$$



- IBP matrix entries are sampled as $X_{nk} \sim_{\text{iid}} \text{Bernoulli}(w_k)$.
- Beta process is the de Finetti measure of the IBP, that is, $Q = \text{BP}$.
- θ is a random measure (but not normalized)

REFERENCES I

[FLP12] S. Favaro, A. Lijoi, and I. Prünster. Conditional formulae for Gibbs-type exchangeable random partitions. *Ann. Appl. Probab.* To appear., 2012.

[GG06] T. L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*, volume 18, 2006.

[GG11] T. L. Griffiths and Z. Ghahramani. The Indian buffet process: An introduction and review. *J. Mach. Learn. Res.*, 12:1185–1224, 2011.

[GHP07] A. V. Gnedin, B. Hansen, and J. Pitman. Notes on the occupancy problem with infinitely many boxes: General asymptotics and power laws. *Probability Surveys*, 4:146–171, 2007.

[GP06a] A. Gnedin and J. Pitman. Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, 138(3):5674–5684, 2006.

[GP06b] A. Gnedin and J. Pitman. Exchangeable Gibbs partititions and Stirling triangles. *J. Math. Sci.*, 138(3):5674–5685, 2006.

[Hjo90] N. L. Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.*, 18:1259–1294, 1990.

[IJ01] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[JLP09] L. F. James, A. Lijoi, and I. Prüenster. Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36:76–97, 2009.

[Kal05] Olav Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Springer, 2005.

[Kin75] J. F. C. Kingman. Random discrete distributions. *Journal of the Royal Statistical Society*, 37:1–22, 1975.

[LMP05a] A. Lijoi, R. H. Mena, and I. Prüenster. Hierarchical mixture modelling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association*, 100:1278–1291, 2005.

[LMP05b] A. Lijoi, R. H. Mena, and I. Prünster. Hierarchical mixture modeling with normalized inverse-Gaussian priors. *J. Amer. Statist. Assoc.*, 100:1278–1291, 2005.

[LP10] A. Lijoi and I. Prünster. Models beyond the Dirichlet process. In N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors, *Bayesian Nonparametrics*. Cambrdige University Press, 2010.

[Nea00] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.

REFERENCES II

[Pem07] R. Pemantle. A survey of random processes with reinforcement. *Probab. Surv.*, 4:1–79, 2007.

[Pit03] J. Pitman. Poisson-Kingman partitions. In D. R. Goldstein, editor, *Statistics and Science: a Festschrift for Terry Speed*, pages 1–34. Institute of Mathematical Statistics, 2003.

[Rob95] C. P. Robert. Mixtures of distributions: inference and estimation. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*. Chapman & Hall, 1995.

[Sch95] M. J. Schervish. *Theory of Statistics*. Springer, 1995.

[Teh06] Y. W. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992, 2006.

[TJ07] R. Thibaux and M. I. Jordan. Hierarchical beta processes and the Indian buffet process. In *J. Mach. Learn. Res. Proceedings (AISTATS)*, volume 2, pages 564–571, 2007.